Cell PRESS
Open ACCESS
Cell Reports
Article

RNA-DNA Differences Are Generated in Human Cells within Seconds after RNA Exits Polymerase II

Isabel X. Wang,[1,9] Leighton J. Core,[2,9] Hojoong Kwak,[2,4] Lauren Brady,[3] Alan Bruzel,[1,4] Lee McDaniel,[5] Allison L. Richards,[6] Ming Wu,[4] Christopher Grunseich,[7] John T. Lis,[2,*] and Vivian G. Cheung[1,4,8,*]
[1]Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
[3]Cell and Molecular Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[5]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[6]Human Genetics Graduate Program, University of Michigan, Ann Arbor, MI 48109, USA
[7]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA
[8]Departments of Pediatrics and Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[9]These authors contributed equally to this work
*Correspondence: johnlis@cornell.edu (J.T.L.), vgcheung@umich.edu (V.G.C.)
http://dx.doi.org/10.1016/j.celrep.2014.01.037

SUMMARY

RNA sequences are expected to be identical to their corresponding DNA sequences. Here, we found all 12 types of RNA-DNA sequence differences (RDDs) in nascent RNA. Our results show that RDDs begin to occur in RNA chains ~55 nt from the RNA polymerase II (Pol II) active site. These RDDs occur so soon after transcription that they are incompatible with known deaminase-mediated RNA-editing mechanisms. Moreover, the 55 nt delay in appearance indicates that they do not arise during RNA synthesis by Pol II or as a direct consequence of modified base incorporation. Preliminary data suggest that RDD and R-loop formations may be coupled. These findings identify sequence substitution as an early step in cotranscriptional RNA processing.

INTRODUCTION

DNA carries instructions for cellular proteins by providing the code that is transcribed into mRNA that in turn is translated into proteins. It is generally assumed that DNA sequences are copied faithfully into RNA. However, there are exceptions to this one-to-one relationship between RNA and its corresponding DNA sequences. The first example of a transcript sequence that is not encoded by DNA was reported by Benne et al. (1986), who showed that the *coxII* mRNA in trypanosome has four nucleotides that are not encoded in the DNA. They then coined the term "RNA editing" for this "novel mechanism of gene expression" (Benne et al., 1986). Other examples of RNA editing were soon discovered in organisms ranging from plants to metazoans (Cattaneo et al., 1989; Driscoll et al., 1989; Gott et al., 1993; Gualberto et al., 1989). In humans, RNA editing occurs in processes mediated by the ADAR (adenosine deaminases that act on RNA) (Bass and Weintraub, 1988) and APOBEC (apolipoprotein B mRNA editing enzymes) (Chen et al., 1987; Powell et al., 1987) families of proteins. This leads to A-to-G (adenosine to inosine, which is then recognized as guanosine) and C-to-U (cytidine to uridine) changes. Recent advances in sequencing technologies have enabled deep sequencing of DNA and RNA, allowing other investigators (Alon et al., 2012; Bar-Yaacov et al., 2013; Chen, 2013; Chen et al., 2012; Ju et al., 2011; Lagarrigue et al., 2013; Peng et al., 2012; Silberberg et al., 2012; Vesely et al., 2012) and us (Li et al., 2011) to uncover more RNA-DNA sequence differences (RDDs) than canonical RNA-editing events. In different human cells and using various sequencing and analytical methods, we and others have found all 12 types of RDDs.

Although the mechanisms that mediate A-to-G and C-to-U editing in humans are known, we do not know how the other types of RDDs arise. For instance, A-to-C transversions are not likely to be mediated by the ADAR and APOBEC families of deaminases. In order to distinguish among the different types of underlying mechanisms, in this project we sought to determine when RDDs arise. For this purpose, we compared nascent RNA sequences with their corresponding DNA sequences. The results show that all 12 types of RDDs occurred early during transcription. We found RDDs in transcripts beginning at ~55 bases from the active site or ~35 bases beyond the exit channel of RNA polymerase II (Pol II). This demonstrates that the RDD events occur by a mechanism that is distinct from altered base selectivity during catalysis of chain elongation by Pol II. Nonetheless, the RNA processing events that mediate RDDs are closely coupled temporally and spatially to transcription in human cells. Given that RDDs emerge so soon after transcription, we studied

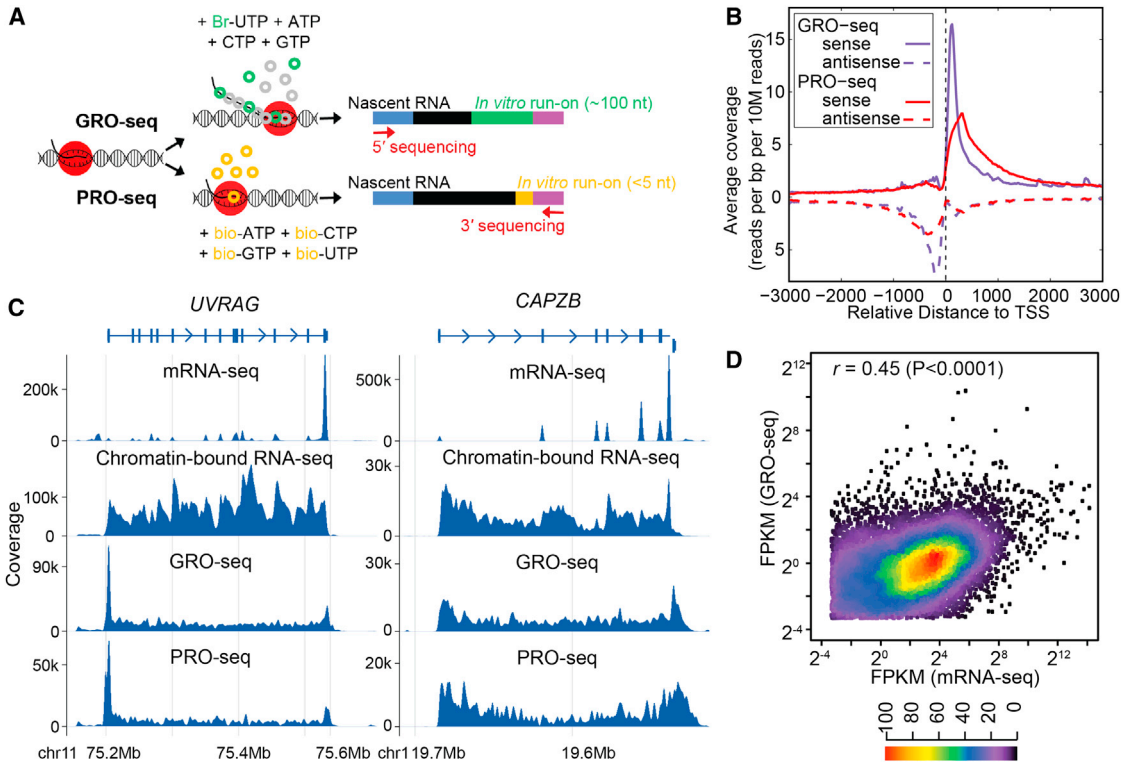

Figure 1. GRO-Seq and PRO-Seq Analysis

(A) Schematic of GRO-seq and PRO-seq.

(B) Comparison between GRO-seq and PRO-seq. Sense and antisense transcripts associated with TSSs are shown for GRO-seq and PRO-seq samples. A slight shift of the PRO-seq promoter-proximal peak downstream relative to the GRO-seq peak is seen because the PRO-seq reads that were <35 nt were not mapped in the analysis, and because GRO-seq maps 5′ ends and PRO-seq maps 3′ ends of nascent RNAs.

(C) mRNA-seq, chromatin-bound nascent RNA-seq, GRO-seq, and PRO-seq results for two representative genes, *UVRAG* and *CAPZB*. For genes with proximal Pol II pausing, such as *UVRAG*, there are more reads mapping to the 5′ ends of genes in both GRO-seq and PRO-seq samples. The schematic gene structure is aligned to mRNA-seq results, with boxes representing exons, lines representing introns, and arrowheads showing the direction of transcription. Coverage is calculated using a bin size of ~1,500 bp and 600 bp, respectively.

(D) Scatterplot of gene expression levels from GRO-seq and mRNA-seq (FPKM > 0.1). Results from GM12750 (shown) and GM12004 are similar (r = 0.45 for both samples). The heatmap indicates the frequency of different expression levels.

cells from a patient with an autosomal-dominant form of juvenile amyotrophic lateral sclerosis (ALS) due to a mutation in the senataxin gene, and found suggestive evidence that RDD formation may be coupled to R loops.

RESULTS

Nascent RNA from GRO-Seq and PRO-Seq

To determine whether RDDs occur during or after transcription, we sequenced nascent RNA using two global run-on sequencing methods: GRO-seq (Core et al., 2008) and precision run-on sequencing (PRO-seq; Figure 1A; Kwak et al., 2013). We obtained ~100 million 100 nt uniquely mapped GRO-seq reads from B cells of two individuals. For one subject, we carried out two independent PRO-seq experiments and obtained ~60 million uniquely mapped reads in each experiment. Additionally, we isolated and sequenced nascent RNA with an alternate method (Wuarin and Schibler, 1994) for comparison (chromatin-bound RNA-seq; ~190 million uniquely mapped reads). Finally, we carried out mRNA sequencing (mRNA-seq) and obtained ~135 million uniquely mapped RNA-seq reads, and sequenced the corresponding genomic DNA of the two individuals to 30× and 60× coverage.

We began by assessing the distributions of mapped reads from the libraries obtained by these four independent methods. As expected (Core et al., 2008; Kwak et al., 2013), GRO-seq and PRO-seq enriched for sequences near transcription start sites (TSSs; Figure 1B). This enrichment in mammalian cells is due to promoter proximal pausing (sense strand) and upstream divergent transcription (antisense strand) (Core et al., 2008; Seila et al., 2008). Additionally, the GRO-seq and PRO-seq data provided sensitive detection of active transcription units and identified >9,000 transcriptionally active genes. To ensure that we were looking at very nascent RNAs, we assessed the extent of splicing in GRO-seq and PRO-seq relative to chromatin-bound nascent RNA and mRNA. Whereas ~20% of the mRNA-seq reads and 5% of the chromatin-bound nascent transcripts covered exon-exon junctions, <1% of the GRO-seq and PRO-seq reads spanned junctions. These nascent transcripts mapped throughout transcription units, including introns (Core and

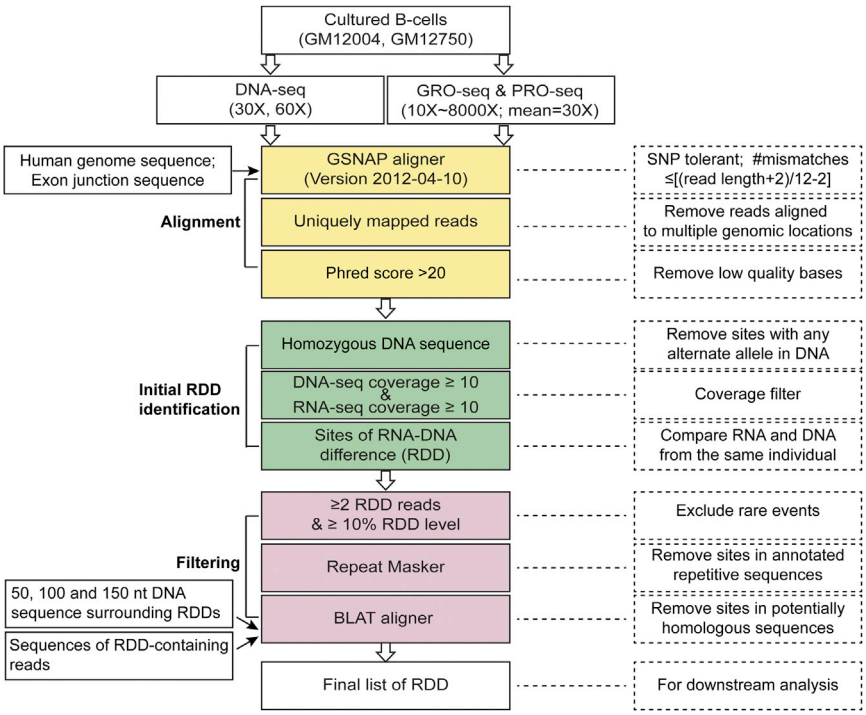

Figure 2. Analysis Steps to Identify RDDs

The potential RDD sites were processed in multiple steps with stringent thresholds to confirm their unique genomic locations. See also Figures S1 and S3, and Tables S1, S2, and S3.

Lis, 2008; Core et al., 2008, 2012), whereas mRNA-seq libraries were dramatically depleted of introns but enriched in the 3′ UTRs due to the sample preparation for polyadenylated transcripts. These findings support the notion that GRO-seq and PRO-seq correspond to greatly enriched short nascent RNA that is newly synthesized (also referred to as "very nascent RNAs" below), while chromatin-bound RNA represents longer transcripts on average from a later stage (referred to as "nascent RNA" below). Figure 1C shows representative results for *UVRAG* and *CAPZB* from sequencing nascent and mature RNAs.

We also compared the expression levels of genes in the very nascent and mature mRNAs. The very nascent RNA differs from mature RNA in that the very nascent RNA levels depend on the density of transcribing Pol II, whereas the mRNA levels depend on the rate of both transcription and mRNA decay. However, the levels of transcripts in the two are significantly correlated ($r = 0.45$, $p \ll 0.0001$; Figure 1D), with outliers representing very stable or unstable mRNAs.

RDDs in Very Nascent RNA

Next, we turned to study RDDs in nascent RNA. Defining when RDDs arise during nascent transcription should help rule out or support particular mechanisms by which they are generated. Therefore, we analyzed the RNA sequences and their corresponding DNA sequences to assess how early during transcription the RDDs arise. The steps taken to identify RDDs are shown in Figure 2. At sites that are covered by at least ten uniquely mapped GRO-seq or PRO-seq reads and ten monomorphic DNA reads (that contain only one nucleotide type [A, C, G, or T]), we compared the nascent RNA and corresponding DNA sequences, and identified sites where RNA and DNA sequences are discordant. For a site to be identified as a candidate RDD, at least 10% of the GRO-seq or PRO-seq reads at that site (and a minimum of two unique reads) must contain a sequence that differs from the underlying DNA sequences. All of the resulting potential RDD sites were further processed in multiple steps to confirm their unique genomic locations.

The results uncovered 2,806 RDDs in one subject (GM12004) and 2,881 RDDs in the other individual (GM12750; Tables S1 and S2). The orientation-specific sequencing allowed us to distinguish all 12 possible types of mismatches between the DNA and the corresponding RNA sequences. In this analysis, we excluded C-to-T RDDs because the use of 5-bromouridine 5′-triphosphate (BrUTP) in GRO-seq may favor this type of misincorporation (Yu et al., 1993). All of the 11 remaining types of RDDs were found, and C-to-G was the most common in both samples (Figure 3A). We analyzed the PRO-seq data in the same way. Except for the 3′-most nucleotide, the sequenced RNAs from the PRO-seq sample were made in the cell (as opposed to about half in GRO-seq), and thus gave us longer segments of in vivo synthesized RNAs for analysis. We found 23,093 RDD sites out of ~115 million nucleotides screened, corresponding to one to two RDD per 10,000 bases screened and a frequency of $\sim 2 \times 10^{-4}$ RDDs in the PRO-seq sample (Table S3), which is comparable to the frequency of RDD in mRNAs (also 10^{-4}) (Li et al., 2011). All 12 types of RDDs were identified (Figure 3B). Even though both GRO-seq and PRO-seq are global run-on assays coupled with deep sequencing, they are not identical; therefore, different numbers of RDDs were detected in the two assays. Unlike GRO-seq, PRO-seq does not use BrUTP and hence misincorporation that favors C-to-T discordance is not a concern; therefore, we included all 12 types of RDDs in our analysis. This added more than 1,700 RDD sites (1,793 C-to-T). In addition, nearly all of the PRO-seq transcripts (except one or at most a few bases), as compared with ~15–20% of the GRO-seq transcripts, are made in vivo. Together, the addition of the C-to-T sites and the longer in vivo synthesized transcripts allowed us to identify ~8-fold more RDD sites in PRO-seq than in GRO-seq. Despite the differences in number, the distributions of RDD types are similar between GRO-seq and PRO-seq samples and across different thresholds of coverage and RDD levels (Figure S1). This reflects the robustness of our analysis. To be certain of our results, we confirmed the mapping and the sequences of the RDD sites in five different experiments and

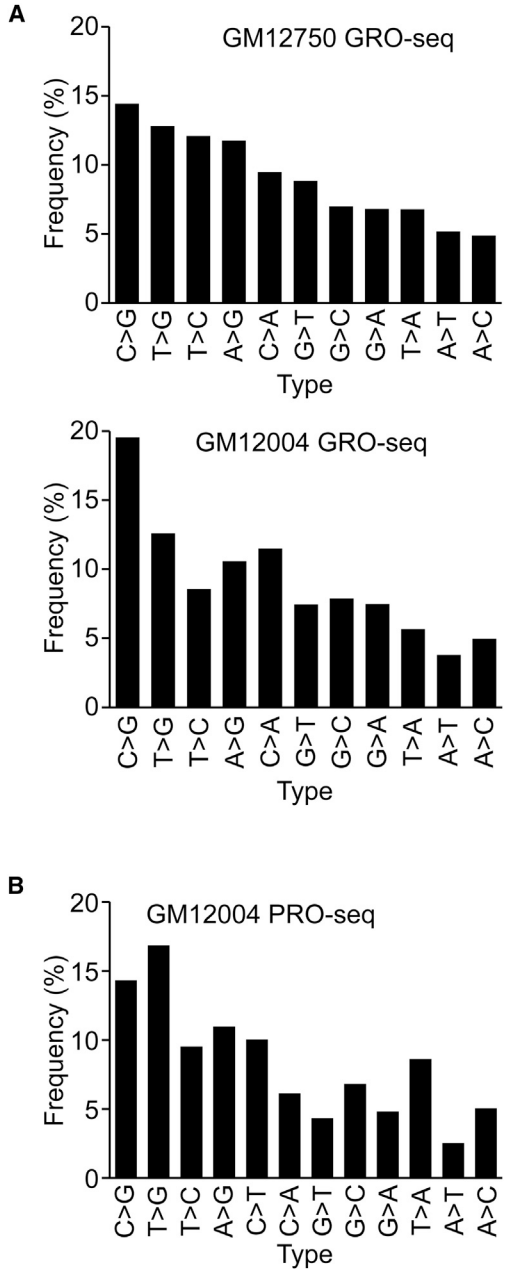

Figure 3. RDDs in Very Nascent Transcripts

(A and B) Distributions of RDD types in (A) GRO-seq samples of two individuals and (B) PRO-seq. RDD types are ordered similarly in (A) and (B), except that in (B), C-to-T RDDs from the PRO-seq sample are included.

See also Figures S2 and S4.

analyses, including genome walking, Sanger sequencing, and droplet digital PCR (ddPCR) using DNA and RNA from multiple tissues (Tables 1 and 2; Figure S2; see Supplemental Results and Discussion for details).

Next, we examined RDDs from different experiments for overlaps. As one would expect, the overlaps of RDD sites between the run-on experiments are low, since the ability to resample an RDD site in independent run-on assays depends on several

Table 1. Results of Genome Walking Confirm that RDDs Are in Unique Regions of the Genome

		5′ Primer		3′ Primer	
Genomic Location	RDD Type	Sequence	No. of Clones	Sequence	No. of Clones
chr1:152175284	G-to-A	G	31	G	9
chr6: 107088915	T-to-A	T	19	T	21
chr9:34336911	G-to-A	G	14	G	10
chr11:72079055	T-to-C	T	10	T	10
chr12:100980077	A-to-C	A	11	A	18
chr14:20221257	G-to-T	G	14	G	16
chr16: 2140620	A-to-G	A	2	A	14
chr19: 2427122	C-to-A	C	15	C	1
chr22:42867269	T-to-C	T	13	T	14
chrX:7004437	G-to-T	G	11	G	10

parameters, including the density of transcribing Pol II, sequence depth, and RDD levels. GRO-seq and PRO-seq identify RDD sites in nascent RNA sequences that are closely associated (<100 nt) with actively transcribing polymerases. Finding the same RDD event in two independent samples relies on sampling an RDD-bearing transcript bound to actively transcribing polymerases in both experiments, and the chance of such an occurrence is very low. The RDD identification also depends on sequence depths and the RDD levels (= number of RDD-containing reads / total number of reads at the site). The median RDD level among the sites detected in GRO-seq and PRO-seq is 0.24; therefore, high coverage (~40×) is needed to obtain 80% of them in replicate samples (Chen, 2013). Nonetheless, 108 RDD sites were found in more than one sample (among the two GRO-seq and one PRO-seq data sets). The RDD sites we found in nascent RNAs were also present at a later stage of transcription. In chromatin-bound transcripts where we have longer transcripts and deeper coverage, we found >1,000 RDD sites from one of the GRO-seq and/or PRO-seq libraries. The distributions of these RDD sites are similar to those observed in GRO-seq and PRO-seq: T-to-G is one of the more abundant types and A-to-T is less frequent. These results show that the RDDs in nascent RNAs can be identified by different assays.

RDD Formation Occurs within Seconds after Transcription

To address how early during transcription RDD events emerge, we first examined the GRO-seq results. As shown in Figure 1A, the GRO-seq reads comprise very nascent RNAs that are transcribed in vivo before nuclei isolation occurs and a portion that are transcribed in vitro during the run-on. Since our very nascent RNAs are triple selected for BrU incorporation and we selectively analyzed reads with an identifiable 3′ end of the nascent RNA, the 3′ portion must contain the in vitro transcribed RNA and the 5′ portion must contain some in vivo synthesized RNA. For both B cell samples, the majority of the RDDs are found in the 5′ portion of the GRO-seq samples, which is enriched for the in vivo made nascent RNA (Figure 4A). These represent newly synthesized transcripts that have just exited the actively

Table 2. ddPCR Validation of GRO-seq RDDs

Genomic Location	Gene Name	RDD Type	Feature	Individual	Level in Nascent RNA	
					GRO-seq (%)[b]	ddPCR (%)
1:152175284	*DENND4B*[a]	G-to-A	coding exon	GM12004	20	15
				GM12750	0	0
3:197100758	*TNK2*	G-to-T	intron	GM12004	0	0
				GM12750	9	9
6:161450890	*MAP3K4*[a]	G-to-T	coding exon	GM12004	17	1
				GM12750	0	2
6:37987903	*ZFAND3*	G-to-C	intron	GM12004	0	0
				GM12750	7	3
9:34336911	—	G-to-A	intergenic	GM12004	8	19
				GM12750	14	27
11:58103493	*ZFP91*	G-to-C	coding exon	GM12004	0	0
				GM12750	18	10
11:72079055	*ARAP1*[a]	T-to-C	intron	GM12004	0	0
				GM12750	11	7
12:100980077	—	A-to-C	intergenic	GM12004	18	17
				GM12750	0	0
16:69880869	*FTSJD1*	C-to-G	5′ UTR	GM12004	9	3
				GM12750	0	0
17:30949447	*AP2B1*	G-to-C	coding exon	GM12004	0	0
				GM12750	10	16
18:8628755	*RAB12*[a]	T-to-C	3′ UTR	GM12004	0	0
				GM12750	11	9
17:34815068	*MED1*	G-to-T	3′ UTR	GM12004	0	0
				GM12750	13	10
19:2197783	*SF3A2*	G-to-C	coding exon	GM12004	0	0
				GM12750	33	10
X:7004437	*HDHD1A*[a]	G-to-T	intron	GM12004	43	50
				GM12750	0	0

[a]Also found in nuclear RNA fractions of both individuals (Figure S4), except that the site in HDHD1A was found in GM12004 but not GM12750.
[b]We included a few RDD sites with levels < 10% in the validations even though the analyses focused on sites with levels > 10%. As shown, even the sites with lower levels were validated by ddPCR analysis of these same libraries.

transcribing polymerase. These findings suggest that RDDs result from transcription-coupled RNA processing steps.

To further refine the time frame for these RDD events, we used PRO-seq to localize more precisely the RDD sites relative to actively transcribing RNA Pol II. In PRO-seq, the in vitro run-on assay was allowed to proceed for only one or at most a few nucleotides; thus, the 3′ ends of the PRO-seq reads mark precisely the locations of the transcriptionally active RNA polymerases in our B cells. This offers an opportunity to examine nascent RNAs that have just exited the active site of Pol II. We examined where the RDDs were found relative to the actively transcribing Pol II, and as seen in the GRO-seq data, the RDD events occurred after the RNA had exited the polymerase (Figure 4B). Moreover, the increased precision and accuracy afforded by PRO-seq allowed us to observe the abrupt increase at ~55 nt from the active site of Pol II, corresponding to the sharp increase in RDD events around position 40 of the PRO-seq reads. As depicted in Figure 4B, the first ~20 bases from the 3′ ends of the reads are nascent RNAs covered by RNA Pol II; thus, RDD sites begin to appear ~35 bases after the RNA exits the polymerase. To confirm this observation, we repeated a PRO-seq experiment. The results confirmed our finding of an increase in RDD at ~55 nt from the active site of Pol II (Figure S3A). In contrast, the RDDs found in mature mRNAs were more uniformly distributed, as expected (Figure 4C). Moreover, analysis of the sequencing quality score shows that the increase in RDD around 55 nt is not a result of a loss of fidelity (Figure S3B). These results are consistent with those obtained from GRO-seq and demonstrate that RDD events appear to occur very rapidly (within seconds) after the nascent RNA is exposed, and do not occur in the Pol II active site during the catalytic step of synthesizing RNA.

RDD Frequency Is Lower in Cells from a Patient with the Senataxin Mutation

Our finding that RDDs emerge soon after nascent transcripts exit from transcription bubbles suggests the coupling of RDDs with

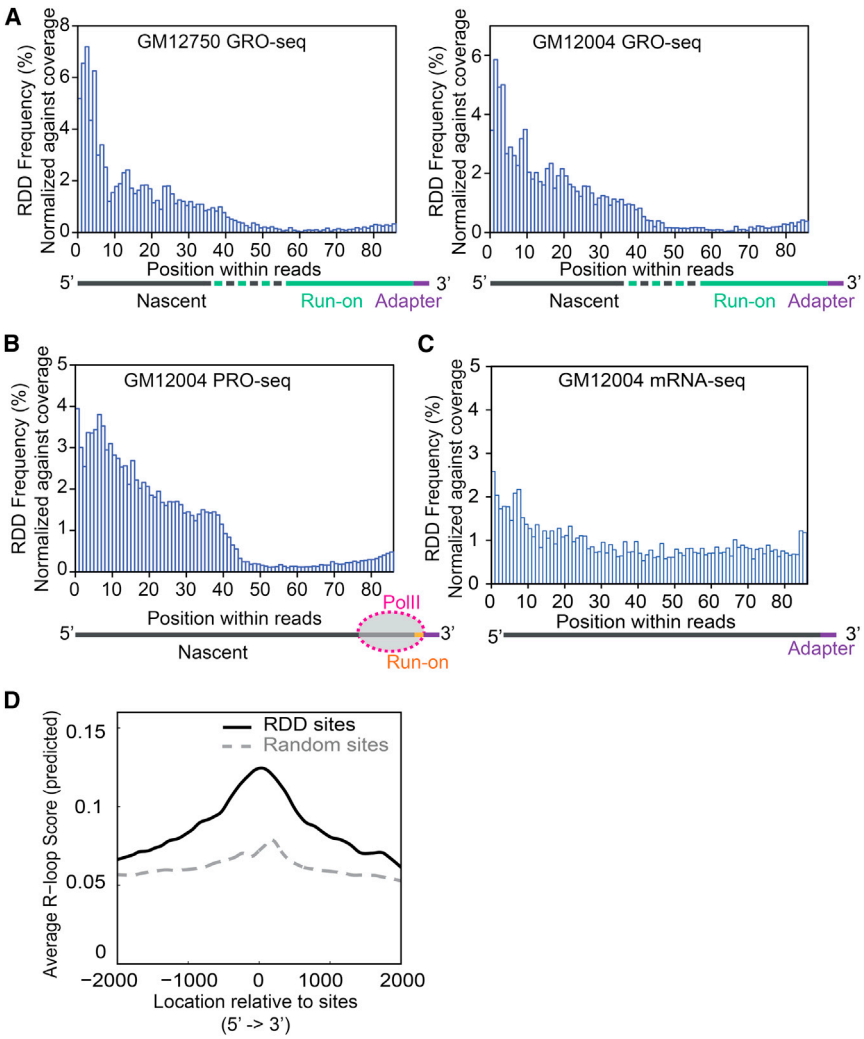

Figure 4. Locations of RDD Sites within Sequencing Reads

(A) Locations of RDD sites along GRO-seq reads. Only reads that have defined 3′ ends (reads that contain 3′ adaptor sequences) were included in our analysis.

(B) Locations of RDD sites along PRO-seq reads.

(A and B) Schematic diagrams indicate the locations of the different segments of GRO-seq (A) and PRO-seq (B) transcripts along the sequence reads.

(C) Locations of RDD sites along mRNA-seq reads.

(D) R-loop-forming sequences are enriched in regions immediately adjacent to RDD sites. Average R-loop scores for 2 kb of regions up and downstream of RDD sites are shown. RDD sites have significantly higher R-loop scores ($p < 0.001$, t test) than random control sites.

See also Figures S2 and S3.

R loops (White and Hogness, 1977), which also initiate behind RNA polymerase. We therefore examined this issue and found that RDDs are enriched significantly ($p < 0.001$) in regions with R-loop-forming sequences (Figure 4D; Ginno et al., 2012; Wongsurawat et al., 2012). To study the co-occurrence of RDDs and R loops, we carried out PRO-seq using cells from a patient with an autosomal-dominant form of juvenile ALS (ALS4) due to a mutation (L389S) in the Senataxin (*SETX*) gene that encodes a DNA/RNA helicase (Chen et al., 2004). The senataxin protein, SETX, interacts with RNA Pol II (Chen et al., 2006; Ursic et al., 2004; Yüce and West, 2013) and plays a role in resolving R loops, particularly in transcription pause sites (Mischo et al., 2011; Skourti-Stathaki et al., 2011; Suraweera et al., 2009; Yüce and West, 2013). The mutation at position 389 corresponds to the N terminus of SETX, which interacts with other nuclear proteins, including RNA Pol II (Yüce and West, 2013). We found 50% fewer RDDs in the very nascent RNA of the ALS4 sample, a frequency of 9×10^{-5} compared with 2×10^{-4} in normal controls. Compared with controls, the RDD sites in the ALS4 sample skewed away from G-bearing transcripts: there were significantly more ($p = 0.03$; t test) RDD events that converted G in the DNA (coding/non-template strand) to other bases in the RNA (32% versus 12% G-to-X, where X = A, C, or T [U]). Since R loops preferentially form around nascent RNA that is G-rich (Roy and Lieber, 2009), this observation suggests that the fewer RDDs in the ALS4 sample may be due to less efficient resolution of R loops. These results encourage further studies to uncover the mechanistic connection between R loops and RDDs.

A-to-G RDDs in Very Nascent RNA Are Not Mediated by ADAR

In our B cells, the only known editing mechanism is ADAR-mediated A-to-G editing (APOBEC1 is not expressed), so we asked whether the A-to-G discrepancies in the nascent RNAs could be explained by ADAR proteins. Previously, ADAR-mediated editing was found in nascent RNA of *Drosophila* (Rodriguez et al., 2012), where nascent RNA was defined as chromatin-bound transcripts. We examined our chromatin-bound transcripts and mature poly-adenylated RNA, and found A-to-G editing events in both fractions, consistent with results in *Drosophila*. However, we did not find these A-to-G sites in GRO-seq or PRO-seq. For example, from mRNA-seq, we identified 65 A-to-G sites in *POLH*, and 48 of the adenosines were also edited in chromatin-bound RNA; however, none of these A-to-G sites were detected in the nascent RNA from GRO-seq or PRO-seq despite good sequence coverage (Figure S3C). For a more comprehensive analysis, we turned to results from several recent studies that identified over 10,000 A-to-G editing sites (Bahn et al., 2012; Carmi et al., 2011; Kiran and Baranov, 2010; Li et al., 2009; Peng et al., 2012). None of the RDD sites in GRO-seq overlap with the editing sites reported in those studies. However, there are some A-to-G events in nascent RNA from GRO-seq and PRO-seq, so we compared the features of these A-to-G sites in nascent RNA with those known to be edited by ADAR-mediated

Table 3. Genes with RDDs in their Nascent RNAs Are Enriched for Roles in Regulation and Metabolism of Macromolecules

GO Term	Examples	p Value
Gene expression	*RNF10, ZNF791, KDM2B; DHX9; ELF4*	1.8×10^{-60}
Nucleic acid metabolic process	*SP3, MAX, RPS6KA4; PSMD11; UTP23*	6.2×10^{-60}
RNA metabolic process	*RPS24; ELF1; CPEB2; DHX9; NFX1*	2.6×10^{-58}
Cellular macromolecule biosynthetic process	*DPF1; SEC14L2; RPL18A; UPF1; HARS*	4.5×10^{-53}
Macromolecule biosynthetic process	*ARFRP1;CTBP2; TSG101; GTF3C2; PARP10*	4.3×10^{-51}
Regulation of macromolecule metabolic process	*AXIN1; FYN; VCP; SMARCA5; ZNF7*	3.9×10^{-50}
Regulation of cellular metabolic process	*BCOR; ELL; MTF1; STAT5A; VPS36*	2.4×10^{-49}
Cellular protein metabolic process	*CCT8; TCF3; RNF115; UBE4B; LNX1*	4.9×10^{-49}
Regulation of primary metabolic process	*ATG7; CLIP3; YLPM1; CD44; POGK*	8.6×10^{-47}
Regulation of nitrogen compound metabolic process	*AGRN; SMARCC1; MOV10; SUMO1; HSPA8*	4.6×10^{-36}

deamination. We found that the sequence characteristics of the A-to-G sites in nascent and mature RNAs appear to be different. Most (>95%) of the ADAR-mediated A-to-G sites in polyadenylated mRNAs are found in Alu repeats (Athanasiadis et al., 2004; Chen, 2013), but in contrast, the A-to-G sites in very nascent (GRO-seq or PRO-seq) RNAs are not in Alu-containing regions. In addition, the A-to-G sites in very nascent RNAs do not have the sequence motif (5′ depletion of G [Lehmann and Bass, 2000]) that flanks ADAR-edited adenosines (Figure S4A; Wang et al., 2013). These data suggest that there are two distinct classes of A-to-G mismatches: those that are mediated by ADAR and those that occur by a separate mechanism on very nascent RNA during transcription.

Other Characteristics of RDDs in Very Nascent RNA

Previous studies of RDDs focused on polyadenylated mRNAs (Bahn et al., 2012; Ju et al., 2011; Li et al., 2011; Peng et al., 2012). The very nascent RNAs in the present study allowed us to assess RDDs in regions such as introns that were spliced out in mature transcripts. Many of the RDDs in very nascent RNAs are found in intronic regions (28%), which could potentially affect downstream RNA processing steps. In addition, nearly half (44%) of the RDDs are intergenic (many of these correspond to gene isoforms with longer 5′ and 3′ UTRs relative to the RefSeq forms). The remaining RDDs (28%) are found in exonic regions and are evenly divided among coding exons and UTRs (48% and 52%, respectively). As we found previously (Li et al., 2011), unlike SNPs, there is no bias against nonsynonymous changes, since ~70% of the coding RDD sites lead to alternate amino acids as predicted by the codon table. We studied the genes that contain RDD sites in nascent RNA and found that they are significantly ($p < 10^{-30}$) enriched for roles in regulation and metabolism of nucleic acids and other macromolecules (see Table 3).

We also examined the sequences (10 bases) surrounding the RDD sites and found that the sequence context may be important. RDDs with the same DNA base share similar sequence characteristics. In particular, C-to-A and C-to-G, and the G-to-A, G-to-C, and G-to-T RDDs share similar surrounding sequences. The RDDs whose DNA base is C reside in regions that are significantly more C rich, whereas RDDs whose DNA base is G reside in regions that are significantly more G rich than negative controls (Figures S4B and S4C; t test, $p < 0.05$). The enrichments of these nucleotides extend in both the 5′ and 3′ directions. These regions are more C rich and G rich, but they are not homopolymer tracts of Cs or Gs (Figure S4D). Thus, they are different from the cotranscriptional editing of homopolymer tracts in Ebola (Volchkov et al., 1995) and paramyxoviruses (Cattaneo et al., 1989; Paterson and Lamb, 1990). Additionally, RDDs whose DNA base is C show depletion of G at the base 3′ of the RDD, and those whose reference base is G show depletion of C at the base 5′ of the RDD. These features may affect the DNA and/or RNA structures, or possibly an RNA/DNA hybrid, which in turn signals for an RDD event as mentioned above.

DISCUSSION

In this work, we examined where RDDs occur and considered the results in the context of known RNA-editing mechanisms. We showed that all 12 types of RDDs are found in RNAs that have recently extruded from the RNA Pol II exit channel. The RDD events occurred in vivo on transcripts ~35 nt from the exit channel of Pol II. Pol II elongates in mammalian cells at 20–60 bases per second (Ardehali and Lis, 2009). Therefore, the RDD events found ~35 bases from the exit channel must occur very shortly after nascent RNA synthesis. Thus, our results indicate that RDDs are likely to occur within a few seconds of RNA synthesis and before classic RNA-editing events. RNAs synthesized by RNA Pol II are quickly modified: 5′ caps are added as the RNA end exits the Pol II RNA channel (Rasmussen and Lis, 1993), introns are often spliced cotranscriptionally (Carrillo Oesterreich et al., 2010; Vargas et al., 2011), and 3′ ends are cleaved and polyadenylated before Pol II terminates transcription (Osheim et al., 2002). Based on knowledge about cotranscriptional processing events and results from the present study, we suggest that RDDs occur soon after the capping of the transcripts and before splicing.

Our purpose in examining the timing of RDDs was to narrow the search for the underlying mechanisms that mediate its formation. A cotranscriptional event that coincides temporally with RDD formation is the emergence of the R loop (Broccoli et al., 2000; Drolet et al., 1995; Massé and Drolet, 1999). As a preliminary search for an association between RDDs and the R loop, we studied RDDs in very nascent RNA of cells from a

juvenile ALS patient with a mutation in the senataxin gene (Chen et al., 2004). The RNA/DNA helicase senataxin interacts with RNA polymerase and mediates the resolution of R loops. We found that the patient had ~50% fewer RDDs in her nascent RNAs. The RDDs seem to be associated with the R loop since there is enrichment in R-loop-forming sequences (Ginno et al., 2012) around RDD sites and depletion of G-bearing RDD transcripts in the patient. These findings point to a possible coupling of RDDs and R-loop formations, and encourage further studies to uncover the molecular basis.

The GRO-seq and PRO-seq assays allowed us to study very nascent RNA for RDD formation. However, these methods also limited our studies to sequences that are covered by or immediately adjacent (<100 bases) to actively transcribing polymerases. It is possible that other mechanisms, such as ADAR-mediated editing, modify RNA transcripts at a later stage of RNA processing. Although our results show that RDD formation occurs very soon after RNA synthesis, they do not imply that all RDD formations have to occur as early cotranscriptional steps. Additional methods may be needed to identify or rule out the existence of other processing steps that modify RNA sequences. Comparing RNA sequences at different stages of maturity alone will not provide a comprehensive view because the levels of many RDD sites are low (<30%), and therefore the depth of sequencing necessary to conclude that an RDD site is absent in one stage of transcript synthesis but present in subsequent stages is difficult to achieve with current sequencing technologies, given the constraints of error rate and cost. However, technologies to isolate RNA from different subcellular compartments and methods for sequence analysis are rapidly improving. They soon will allow the tracking of individual transcripts through various processing steps and thus facilitate the determination of whether there are additional events that modify RNA sequences. In summary, we have identified sequence modification as an early RNA-processing step, thus adding to the already complex set of events that add diversity to transcriptomes.

EXPERIMENTAL PROCEDURES

Cell Culture

Cultured B cells from two normal individuals in the Centre d'Étude du Polymorphisme Humain database, GM12004 and GM12750, were obtained from Coriell Cell Repositories. Skin fibroblasts from a forearm biopsy of the juvenile ALS patient were collected under NIH protocol 00-N-0043, which was approved by the Intramural Combined Neuroscience Institutional Review Board of the NIH.

DNA-Seq

DNA-seq libraries were prepared and sequenced on a HiSeq instrument (Illumina) to obtain 60× and 30× coverage, respectively.

mRNA-Seq and Chromatin-Bound Nascent RNA-Seq

For mRNA sequencing, RNA-seq libraries were prepared according to the Illumina TruSeq RNA sample preparation protocol. Chromatin-bound nascent RNA was extracted as previously described (Wuarin and Schibler, 1994). The mRNA and chromatin RNA were sequenced on a HiSeq instrument.

GRO-Seq and PRO-Seq

Nuclei were isolated from cultured B cells and GRO-seq libraries were prepared with 5 × 10^6 nuclei as described previously (Core et al., 2008, 2012). PRO-seq libraries were prepared as described previously (Kwak et al., 2013). Briefly, 5 × 10^6 nuclei were added to a 2× nuclear run-on reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% sarkosyl, 5 mM $MgCl_2$, 1 mM dithiothreitol, 0.375 mM each of biotin-11-A/C/G/UTP [Perkin-Elmer], 0.8 U/μl RNase inhibitor) and incubated for 3 min at 30°C. Nascent RNA was extracted and fragmented by base hydrolysis in 0.2 N NaOH on ice for 10–12 min, and neutralized by adding a 1× volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was purified using streptavidin beads and ligated with reverse 3′ RNA adaptor (5′p-GAUCGUCGGACUG-UAGAACUCUGAAC-/3′InvdT/), and biotin-labeled products were enriched by another round of streptavidin bead binding and extraction. For 5′ end repair, the RNA products were successively treated with tobacco acid pyrophosphatase (Epicenter) and polynucleotide kinase (NEB). 5′ repaired RNA was ligated to reverse 5′ RNA adaptor (5′-CUGAACAAGCAGAAGACGGCAUACGA-3′) before being further purified by the third round of streptavidin bead binding and extraction. RNA was reverse transcribed using 25 pmol of RT primer (5′AATGATACGGC GACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3′). The product was amplified 15 ± 3 cycles and products >150 bp (insert > 70 bp) were PAGE purified before being analyzed on an Illumina HiSeq 2500 instrument. Two PRO-seq experiments, one at the Lis lab at Cornell University (Figures 3 and 4) and one at the Cheung lab at the University of Pennsylvania, were carried out (Figure S3A).

Sequence Analysis

DNA-seq and RNA-seq reads were aligned to the human reference genome (hg18) using GSNAP (Wu and Nacu, 2010) (version 2012-04-10). A list of SNP sites in the CEU population from Hapmap (release #28) and 1000 Genomes (pilot project) was used for SNP-tolerant alignments. Alignments with (read length + 2)/12 − 2 or fewer mismatches were obtained for each read. PRO-seq sequences were converted to the reverse complements before alignment. For RNA sequence analysis, known exon-exon junctions (defined by RefSeq [downloaded March 7, 2011] and Gencode [version 3c]) and novel junctions (defined by GSNAP) were accepted. Read coverage was analyzed using RSeQC, and the RPKM (read per kilobase per million reads) for each gene was calculated (Wang et al., 2012). For GRO-seq and PRO-seq, we included all of the reads that covered the exon or intron region in computing the RPKM, but excluded the 1 kb region downstream of TSS, which is overrepresented by short transcripts associated with proximally paused Pol II.

RDDs

To identify RDDs, we compared an RNA sequence with its corresponding DNA sequence. Low-quality bases (Phred quality score < 20) in both the RNA and DNA were removed. To be included as RDD sites in the final lists, the following criteria had to be met: (1) a minimum of ten total DNA-seq reads cover the site; (2) the DNA sequence at this site is 100% concordant, without any DNA-seq reads containing alternative alleles; (3) a minimum of ten total RNA-seq reads cover the site; and (4) the level of RDD (number of RNA-seq reads containing non-DNA allele / number of all RNA-seq reads covering a given site) is ≥10% (a minimum of two RNA-seq reads containing RDD). To ensure the accuracy of the RDD sites, we performed additional filtering steps using two additional mapping algorithms. See the Supplemental Experimental Procedures for further details.

ACCESSION NUMBERS

The sequence data have been deposited in the National Center for Biotechnology Information database under accession numbers GSE38233 (mRNA-seq), GSE39878 (chromatin-bound RNA-seq, GRO-seq, and PRO-seq), and ERP001478 (DNA-seq).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results and Discussion, Supplemental Experimental Procedures, four figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.01.037.

ACKNOWLEDGMENTS

We dedicate this paper to the memory of Dr. Tom Kadesch, who suggested the collaboration between the Lis and Cheung labs to study mechanisms that underlie RDDs.

We thank Dr. Nancy Zhang for discussions about estimating sequencing errors, Jonathan Toung for analysis of the sequencing data, and Zhengwei Zhu for data analysis. We thank Dr. Kenneth Fischbeck for discussion about juvenile ALS and patient samples. Part of this work was carried out at the University of Pennsylvania prior to the Cheung lab's move to the University of Michigan. This work was supported by grants from the National Institutes of Health (MH087384 and ES015733 to V.G.C., and GM25232 to J.T.L.) and funds from the Howard Hughes Medical Institute (to V.G.C.).

Received: October 20, 2013
Revised: December 27, 2013
Accepted: January 28, 2014
Published: February 20, 2014

REFERENCES

Alon, S., Mor, E., Vigneault, F., Church, G.M., Locatelli, F., Galeano, F., Gallo, A., Shomron, N., and Eisenberg, E. (2012). Systematic identification of edited microRNAs in the human brain. Genome Res. *22*, 1533–1540.

Ardehali, M.B., and Lis, J.T. (2009). Tracking rates of transcription and splicing in vivo. Nat. Struct. Mol. Biol. *16*, 1123–1124.

Athanasiadis, A., Rich, A., and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. *2*, e391.

Bahn, J., Lee, J., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res. *22*, 142–150.

Bar-Yaacov, D., Avital, G., Levin, L., Richards, A.L., Hachen, N., Rebolledo Jaramillo, B., Nekrutenko, A., Zarivach, R., and Mishmar, D. (2013). RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res. *23*, 1789–1796.

Bass, B.L., and Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell *55*, 1089–1098.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell *46*, 819–826.

Broccoli, S., Phoenix, P., and Drolet, M. (2000). Isolation of the topB gene encoding DNA topoisomerase III as a multicopy suppressor of topA null mutations in Escherichia coli. Mol. Microbiol. *35*, 58–68.

Carmi, S., Borukhov, I., and Levanon, E.Y. (2011). Identification of widespread ultra-edited human RNAs. PLoS Genet. *7*, e1002317.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol. Cell *40*, 571–581.

Cattaneo, R., Kaelin, K., Baczko, K., and Billeter, M.A. (1989). Measles virus editing provides an additional cysteine-rich protein. Cell *56*, 759–764.

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. Proc. Natl. Acad. Sci. USA *110*, E2741–E2747.

Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., Silberman, S.R., Cai, S.J., Deslypere, J.P., Rosseneu, M., et al. (1987). Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science *238*, 363–366.

Chen, Y.Z., Bennett, C.L., Huynh, H.M., Blair, I.P., Puls, I., Irobi, J., Dierick, I., Abel, A., Kennerson, M.L., Rabin, B.A., et al. (2004). DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). Am. J. Hum. Genet. *74*, 1128–1135.

Chen, Y.Z., Hashemi, S.H., Anderson, S.K., Huang, Y., Moreira, M.C., Lynch, D.R., Glass, I.A., Chance, P.F., and Bennett, C.L. (2006). Senataxin, the yeast Sen1p orthologue: characterization of a unique protein in which recessive mutations cause ataxia and dominant mutations cause motor neuron disease. Neurobiol. Dis. *23*, 97–108.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Chen, R., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293–1307.

Core, L.J., and Lis, J.T. (2008). Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science *319*, 1791–1792.

Core, L., Waterfall, J., and Lis, J. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

Driscoll, D.M., Wynne, J.K., Wallis, S.C., and Scott, J. (1989). An in vitro system for the editing of apolipoprotein B mRNA. Cell *58*, 519–525.

Drolet, M., Phoenix, P., Menzel, R., Massé, E., Liu, L.F., and Crouch, R.J. (1995). Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. Proc. Natl. Acad. Sci. USA *92*, 3526–3530.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol. Cell *45*, 814–825.

Gott, J.M., Visomirski, L.M., and Hunter, J.L. (1993). Substitutional and insertional RNA editing of the cytochrome c oxidase subunit 1 mRNA of Physarum polycephalum. J. Biol. Chem. *268*, 25483–25486.

Gualberto, J.M., Lamattina, L., Bonnard, G., Weil, J.H., and Grienenberger, J.M. (1989). RNA editing in wheat mitochondria results in the conservation of protein sequences. Nature *341*, 660–662.

Ju, Y.S., Kim, J.I., Kim, S., Hong, D., Park, H., Shin, J.Y., Lee, S., Lee, W.C., Kim, S., Yu, S.B., et al. (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat. Genet. *43*, 745–752.

Kiran, A., and Baranov, P.V. (2010). DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics *26*, 1772–1776.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Lagarrigue, S., Hormozdiari, F., Martin, L.J., Lecerf, F., Hasin, Y., Rau, C., Hagopian, R., Xiao, Y., Yan, J., Drake, T.A., et al. (2013). Limited RNA editing in exons of mouse liver and adipose. Genetics *193*, 1107–1115.

Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry *39*, 12875–12884.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science *324*, 1210–1213.

Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M., and Cheung, V.G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. Science *333*, 53–58.

Massé, E., and Drolet, M. (1999). Escherichia coli DNA topoisomerase I inhibits R-loop formation by relaxing transcription-induced negative supercoiling. J. Biol. Chem. *274*, 16659–16664.

Mischo, H.E., Gómez-González, B., Grzechnik, P., Rondón, A.G., Wei, W., Steinmetz, L., Aguilera, A., and Proudfoot, N.J. (2011). Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol. Cell *41*, 21–32.

Osheim, Y.N., Sikes, M.L., and Beyer, A.L. (2002). EM visualization of Pol II genes in Drosophila: most genes terminate without prior 3′ end cleavage of nascent transcripts. Chromosoma *111*, 1–12.

Paterson, R.G., and Lamb, R.A. (1990). RNA editing by G-nucleotide insertion in mumps virus P-gene mRNA transcripts. J. Virol. *64*, 4137–4145.

Peng, Z., Cheng, Y., Tan, B.C., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat. Biotechnol. *30*, 253–260.

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J., and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell *50*, 831–840.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc. Natl. Acad. Sci. USA *90*, 7923–7927.

Rodriguez, J., Menet, J.S., and Rosbash, M. (2012). Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Mol. Cell *47*, 27–37.

Roy, D., and Lieber, M.R. (2009). G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol. Cell. Biol. *29*, 3124–3133.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Silberberg, G., Lundin, D., Navon, R., and Öhman, M. (2012). Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. Hum. Mol. Genet. *21*, 311–321.

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol. Cell *42*, 794–805.

Suraweera, A., Lim, Y., Woods, R., Birrell, G.W., Nasim, T., Becherel, O.J., and Lavin, M.F. (2009). Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. Hum. Mol. Genet. *18*, 3384–3396.

Ursic, D., Chinchilla, K., Finkel, J.S., and Culbertson, M.R. (2004). Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. Nucleic Acids Res. *32*, 2441–2452.

Vargas, D.Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S.A., Schedl, P., and Tyagi, S. (2011). Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell *147*, 1054–1065.

Vesely, C., Tauber, S., Sedlazeck, F.J., von Haeseler, A., and Jantsch, M.F. (2012). Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res. *22*, 1468–1476.

Volchkov, V.E., Becker, S., Volchkova, V.A., Ternovoj, V.A., Kotov, A.N., Netesov, S.V., and Klenk, H.D. (1995). GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology *214*, 421–430.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wang, I.X., So, E., Devlin, J.L., Zhao, Y., Wu, M., and Cheung, V.G. (2013). ADAR regulates RNA editing, transcript stability, and gene expression. Cell Rep. *5*, 849–860.

White, R.L., and Hogness, D.S. (1977). R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA. Cell *10*, 177–192.

Wongsurawat, T., Jenjaroenpun, P., Kwoh, C.K., and Kuznetsov, V. (2012). Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. Nucleic Acids Res. *40*, e16.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol. Cell. Biol. *14*, 7219–7225.

Yu, H., Eritja, R., Bloom, L.B., and Goodman, M.F. (1993). Ionization of bromouracil and fluorouracil stimulates base mispairing frequencies with guanine. J. Biol. Chem. *268*, 15935–15943.

Yüce, O., and West, S.C. (2013). Senataxin, defective in the neurodegenerative disorder ataxia with oculomotor apraxia 2, lies at the interface of transcription and the DNA damage response. Mol. Cell. Biol. *33*, 406–417.